The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity

(Team Seekers: Anvitha Karanam, Jahnavi Rangu, Leela Alekhya Vedula, Manisha Yacham)

Procedure followed in the paper:

1. Performed cross domain experiments on data with labeled according to the source reputation and received poor test results
2. Performed batch training on different sets of training data that is reliably labeled according to the veracity of the data from various news articles and received much better test results

Hence the system performance on different test datasets depends on topic distribution and also conclude that collecting well-balanced and carefully-assessed training data is important for developing robust misinformation detection systems.

Datasets:

FEVER (Thorne et al., 2018) dataset contains both claims and texts from Wikipedia pages that support or refute those claims. Our objective is to elaborate on the distinction between classifying reputation-based labeled news articles and individually-assessed news articles. The below is the list of 4 datasets used for misinformation detection.

1. **Rashkin et al. dataset:** This dataset is classified into four classes: propaganda, satire, hoax and trusted. They suspect that the noisy strategy to label all articles of a publisher based on its limits its power to distinguish individual misinformative from truthful news articles and reputation highly biases the classifier decisions.
2. **Rubin et al. dataset:** This dataset contains balanced numbers of individually evaluated satirical and legitimate texts.
3. **BuzzfeedUSE dataset:** It is the first source of information that is used in this procedure to harvest full news articles with veracity labels is from the Buzzfeed fact checking company. The links were collected from nine Facebook pages. The information in each URL is rated by human experts.
4. **Snopes312 dataset:** This dataset is used to harvest full news articles with veracity labels is Snopes, a well-known rumor debunking website run by a team of expert editors.

Summary of Experiments:

The paper suggests that Convolutional Neural Networks (CNNs) and TFIDF complete with each other. The authors have trained various architectures of CNN and classic classifiers like Naive Bayes and SVM with TF-IDF features on Rashkin dataset. The best results to identify misinformation was obtained from SVM classifier using unigram TF-IDF features with L2 regularization. The model performed well on collected items within Rashkin et al.'s test dataset.

However,the performance dropped alot when the model is applied to Rubin et al.'s data. This dataset has balance between the topics of the legitimate instances and satirical instances. This suggests that the classifier depends on the topics of the news articles.

After Rubin et al.'s data, the next experiment was conducted with the same model using BuzzfeedUSE and Snopes312, as test datasets. The first observation is, the classifier shows some vulnerability to true against false information in the first test dataset. Most of the BuzzfeedUSE data is predicted as Hoax, which suggests that the classifier has seen most of the data in BuzzfeedUSE as Hoax in Rashkin's data. Finally, Snopes312 has shown the best results by the classifier with more match between the actual and predicted values compared to the BuzzfeedUSE data. From the above-conducted experiments, the author concludes that the performance of the model alters for various test data and was not able to distinguish between the 4 labels properly. Hence, reputation-based classification is not enough for predicting the veracity level of various news articles.